Nonparametric estimation of mean and dispersion functions in extended generalized linear models

I. Gijbels[1,3], I. Prosdocimi[1,3] and G. Claeskens[2,3]

[1] Department of Mathematics, Katholieke Universiteit Leuven
[2] Operations Research and Business Statistics, Katholieke Universiteit Leuven
[3] Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven, Belgium.

Abstract

We study joint nonparametric estimators of the mean and the dispersion function in extended double exponential family models. The starting point is the exponential family and the generalized linear models setting. The extended models allow for both overdispersion and underdispersion, or even a combination of both. We simultaneously estimate the dispersion function and the mean function by using P-splines with a difference type of penalty to avoid overfitting. Special attention is given to the smoothing parameter selection as well as to implementation issues. The performance of the method is investigated via simulations. A comparison with other available methods is made. We provide applications to several sets of data, including continuous data, counts and proportions.

Keywords and phrases: Double exponential family; Extended quasi-likelihood; Nonparametric regression; Overdispersion; P-splines; Variance estimation; Underdispersion.

1 Introduction

The aim of this work is to provide tools to nonparametrically estimate mean and dispersion functions. In a specific generalized linear model a fixed relationship exists between the mean and the variance. For example, for the Poisson distribution the mean is equal to the variance. Real data often show a larger variability than what we would expect from the theoretical model. This phenomenon is referred to as overdispersion. The opposite situation of less variability occurs less frequently and is called underdispersion. Our approach can deal with both overdispersed and underdispersed data, coming from different distributions originating from the exponential family. As such it is a unified approach not requesting special models for overdispersion such as a beta-binomial for proportion data or a negative binomial for count data. We start from a double exponential family (Efron

(1986)), and model both the mean and the dispersion function as smooth functions of the covariate through P-splines (Eilers and Marx (1996)). The flexible modeling even allows for dispersion functions that might reveal underdispersion in one region of the covariate-domain and overdispersion in other regions.

Hinde and Demétrio (1998) give an overview of estimation methods for overdispersed data. Parametric models for the dispersion as a function of one or more covariates include extended quasi-likelihood (Nelder and Pregibon (1987); McCullagh and Nelder (1989)), the double exponential regression models (see Efron (1986); Galfand and Dalal (1990); Dey *et al.* (1997)), and the pseudo-likelihood approach proposed by Davidian and Carroll (1987), among others. All these approaches have their advantages and disadvantages. See for example Nelder and Lee (1992), Lee and Nelder (2000) and Davidian and Carroll (1988) for a comparison of methods. They all share the common feature of being based on parametric modeling. Some work has been done on flexible modeling of overdispersion, see for example Aerts and Claeskens (1997), Chapter 14 in Ruppert *et al.* (2003) and Nott (2006). The latter uses a mixed model representation of splines in the double exponential family framework to get to a Bayesian estimation of both the mean and the variance function. Rigby and Stasinopoulos (2005) propose generalized additive models for location, scale and shape, in which they nonparametrically estimate different parameters of a distribution using a mixed model approach and concepts of hierarchical modeling. Earlier references on hierarchical modeling include Lee and Nelder (1996, 2001).

Estimating the unknown variance/dispersion is of interest to improve the estimation of the mean and can, in different contexts, be an issue of importance on its own. See for example Carroll and Ruppert (1988), for the importance and utility of estimating variance functions in the linear regression context.

In Section 2 we briefly discuss the framework for the estimation procedure. The nonparametric procedure for estimating mean and dispersion function using penalized regression techniques is presented in Section 3 including a discussion on practical choices of smoothing parameters. A simulation study investigating the performance of the estimation procedure is provided in Section 4, which also includes a comparison with other existing methods. Section 5 contains illustrations with real data examples. In Section 6 we provide some further discussions.

2 The double exponential family for joint modeling of mean and dispersion function

Efron (1986) introduced the double exponential family of distributions. This is a generalization of the exponential family in which an extra parameter controlling the variance is added.

In a one-parameter exponential family, denoted by $(Y|X = x) \sim \text{EF}(b(\theta(x)), \phi)$, the conditional density of the response variable Y given the covariate $X = x$, is of the form

$$e_Y(y; \theta(x), \phi) = \exp\left\{\frac{y\theta(x) - b(\theta(x))}{\phi} + c(y, \phi)\right\}. \tag{2.1}$$

Here $b(\cdot)$ and $c(\cdot)$ are known functions, identifying specific distributions (including the normal, the Poisson and the binomial distribution). The natural location parameter is θ and ϕ is the scale parameter. The moments for the exponential family correspond to $\mu(x) = E[Y|X = x] = b'(\theta(x))$ and $\text{Var}[Y|X = x] = \phi b''(\theta(x))$. It is assumed throughout that $(b')^{-1}$ exists and that $b''(\cdot) \neq 0$.

The double exponential family

For simplicity of presentation we introduce this family in a non-regression setup, as in Efron (1986). The density of a variable Y coming from the double exponential family $Y \sim \text{DEF}(b(\theta), \phi, \gamma))$ is

$$\tilde{f}_Y(y; \theta, \phi, \gamma) = c(\theta, \gamma)\gamma^{-\frac{1}{2}} e_Y(y; \theta, \phi)^{\frac{1}{\gamma}} e_Y(y; \theta_S, \phi)^{1-\frac{1}{\gamma}}, \tag{2.2}$$

where θ_S is the choice of θ corresponding to the saturated one-parameter exponential model $Y \sim \text{EF}(b(\theta), \phi)$, namely $\theta_S = (b')^{-1}(y)$, which maximizes $e_Y(y; \theta, \phi)$ over all possible values of θ and $c(\theta, \gamma)$ is a normalizing constant, such that $\int_{-\infty}^{\infty} \tilde{f}_Y(y; \theta, \phi, \gamma)dy = 1$. Efron (1986) shows that the constant $c(\theta, \gamma)$ in (2.2) can be (first order) approximated by 1, so that an approximation of (2.2) is

$$f_Y(y; \theta, \phi, \gamma) = \gamma^{-\frac{1}{2}} e_Y(y; \theta, \phi)^{\frac{1}{\gamma}} e_Y(y; \theta_S, \phi)^{1-\frac{1}{\gamma}}. \tag{2.3}$$

This approximation works quite well overall, but might be poorer for extremely overdispersed and underdispersed cases with a small mean. For a detailed study on the use of the unnormalized form (2.3) of the density for inference see Lee and Nelder (2000). As suggested and motivated in Efron (1986) we use this approximated formulation of the

double exponential family, which leads to a convenient implementation of the inference procedure.

It can be shown that for $Y \sim \text{DEF}(b(\theta), \phi, \gamma)$ we approximately have that $E(Y) = \mu = b'(\theta)$ and $\text{Var}[Y] = \gamma\phi b''(\theta)$ (see Formula (3.21) in Efron (1986)), and clearly when $\gamma = 1$ the double exponential family reduces to $Y \sim \text{EF}(b(\theta), \phi)$ in (2.1), while for $\gamma > 1 (< 1)$ we have overdispersion (underdispersion). In other words, γ is the parameter describing the dispersion.

For the one-parameter exponential family (2.1), the deviance is defined as

$$d(y, \theta) = 2[\log(e_Y(y; \theta_S, \phi)) - \log(e_Y(y; \theta, \phi))] = 2\left(\frac{y\theta_S - b(\theta_S)}{\phi} - \frac{y\theta - b(\theta)}{\phi}\right). \quad (2.4)$$

Therefore we can rewrite (2.3) as

$$f_Y(y; \theta, \phi, \gamma) = \gamma^{-\frac{1}{2}} \left\{\exp\left[\frac{1}{2}d(y, \theta)\right]\right\}^{-\frac{1}{\gamma}} e_Y(y; \theta_S, \phi).$$

This is helpful in getting an elegant expression of the log-likelihood when considering the regression case.

When using double exponential families in a regression context one is interested in studying how the expected value and the dispersion of Y depend on a covariate X. It is assumed that $(Y|X = x) \sim \text{DEF}(b(\theta(x)), \phi, \gamma(x))$, with ϕ a known constant, not dependent on X. Similarly as in the one-parameter exponential family we introduce a link function g and consider $\eta(x) = g(\mu(x))$, with $\mu(x) = E[Y|X = x]$. In parametric models we might fit a linear function of the covariate, $\eta(x) = \beta_0 + \beta_1 x$, with $\boldsymbol{\beta} = (\beta_0\ , \beta_1)^T$ the vector of unknown parameters. Here the superscript T denotes the transposed of a vector (or matrix). For the canonical link $\eta(x) = g(b'(\theta(x))) = \theta(x)$. From now on, unless otherwise mentioned, we will always use canonical link functions.

When working within the double exponential family we also use a regression model for the dispersion function. In a parametric (linear) setting this would mean that we model a transformation of the dispersion function as a linear combination of the covariate: $\gamma(x) = h(\xi(x))$, $\xi(x) = \delta_0 + \delta_1 x$, with $\boldsymbol{\delta} = (\delta_0\ ,\ \delta_1)^T$ the vector of unknown parameters. Note that the function $h(\cdot)$ needs to be chosen with care since a dispersion function should always be nonnegative ($h \geq 0$). In addition we assume that h is such that h^{-1} exists and $h'(\cdot) \neq 0$.

Throughout the remainder of this work we model the mean and the dispersion function in a double exponential family framework. Note that such a direct modeling of the

dispersion function, can handle both overdispersed and underdispersed data or even data where the appearance of overdispersion or underdispersion depends on the range of values of the covariate. The double exponential family provides a unified framework for different types of data (counts, proportions, etc.), where we can directly model the dispersion function $\gamma(x)$ that can take values either larger or smaller than 1.

3 Nonparametric estimation of mean and dispersion

3.1 Estimation procedure

The goal of this work is to obtain nonparametric estimators for both the mean and the dispersion function combining the parametric approach of the double exponential family where $(Y|X = x) \sim \text{DEF}(b(\theta(x)), \phi, \gamma(x))$, as discussed in Section 2, with the nonparametric P-splines technique of Eilers and Marx (1996).

Via $\eta(x) = g(\mu(x))$ and $\xi(x) = h^{-1}(\gamma(x))$ we model each of these functions as a linear combination of B-spline basis functions (possible different sets for each) and use difference type of penalties to prevent overfitting. For the mean function, for a given set of knots $\{\kappa_1, \ldots, \kappa_{k_\mu}\}$, B-spline basis functions of degree p_μ are composed of polynomial pieces of degree p_μ, joined together in an appropriate way at each knot κ_j. This results into a B-spline basis of dimension $K_\mu = p_\mu + 1 + k_\mu$, denoted by $B_{\mu,1}(\cdot), B_{\mu,2}(\cdot), \ldots B_{\mu,K_\mu}(\cdot)$. For details, see de Boor (2001). The unknown function $\eta(\cdot)$ is then modeled as $\eta(x) = \sum_{j=1}^{K_\mu} \alpha_{\mu,j} B_{\mu,j}(x) = \boldsymbol{B}_\mu^T(x)\boldsymbol{\alpha}_\mu$ where we denote $\boldsymbol{B}_\mu(x) = (B_{\mu,1}(x), \ldots, B_{\mu,K_\mu}(x))^T$, a vector of dimension $K_\mu \times 1$, and $\boldsymbol{\alpha}_\mu = (\alpha_{\mu,1}, \ldots, \alpha_{\mu,K_\mu})^T$, the unknown vector of spline coefficients for the mean function. Similarly, the dispersion function $\gamma(\cdot)$ is modeled in a flexible way via a set of knots $\{\nu_1, \ldots, \nu_{k_\gamma}\}$ and B-splines of degree p_γ, leading to $\xi(x) = \boldsymbol{B}_\gamma^T(x)\boldsymbol{\alpha}_\gamma$, where $\boldsymbol{B}_\gamma(x)$ has dimension $K_\gamma \times 1$ with $K_\gamma = p_\gamma + 1 + k_\gamma$, and $\boldsymbol{\alpha}_\gamma = (\alpha_{\gamma,1}, \ldots, \alpha_{\gamma,K_\gamma})^T$.

Large values of K_μ and K_γ might lead to overfitting. This is avoided by introducing penalty terms on the spline coefficients in the log-likelihood function. P-splines regression uses penalties based on higher order differences of the coefficients. For the mean estimation this leads to a penalty term $\sum_{j=m+1}^{K_\mu} (\Delta^m \alpha_{\mu,j})^2$, where m is the order of the difference operator: $\Delta\alpha_{\mu,j} = \alpha_{\mu,j} - \alpha_{\mu,j-1}$, $\Delta^2\alpha_{\mu,j} = \Delta\Delta\alpha_{\mu,j} = \alpha_{\mu,j} - 2\alpha_{\mu,j-1} + \alpha_{\mu,j-2}$ and so on.

We consider relatively large numbers of knots K_μ and K_γ, which cover the range of observed values of X in the sample $(\boldsymbol{x}, \boldsymbol{y})$ from (X, Y). Based on the two sets of B-spline basis functions and the observed sample, we then define the matrix $\boldsymbol{B}_\mu$ of dimension

$n \times K_\mu$, with i-th row $\boldsymbol{B}_{\mu,i}^T = \boldsymbol{B}_\mu(x_i) = (B_{\mu,1}(x_i), \ldots, B_{\mu,K_\mu}(x_i))$. In a similar manner the matrix $\boldsymbol{B}_\gamma$ of dimension $n \times K_\gamma$ is obtained.

Maximum penalized likelihood inference for $\mu(x)$ and $\gamma(x)$ follows from maximization with respect to $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ of the penalized log-likelihood

$$l(\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_\gamma; \boldsymbol{x}, \boldsymbol{y}, \lambda_\mu, \lambda_\gamma) = \sum_{i=1}^{n} \left\{ \log(h(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)) + \frac{1}{h(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)} d(y_i, \boldsymbol{B}_{\mu,i}^T \boldsymbol{\alpha}_\mu) \right\} - \frac{1}{2}\lambda_\mu \boldsymbol{\alpha}_\mu^T \boldsymbol{D}_m^T \boldsymbol{D}_m \boldsymbol{\alpha}_\mu - \frac{1}{2}\lambda_\gamma \boldsymbol{\alpha}_\gamma^T \boldsymbol{D}_\ell^T \boldsymbol{D}_\ell \boldsymbol{\alpha}_\gamma, \tag{3.1}$$

where $\boldsymbol{D}_m$ and $\boldsymbol{D}_\ell$ are the matrix representations of the m-th and ℓ-th order difference operators, and with $\lambda_\mu > 0$ and $\lambda_\gamma > 0$ the two smoothing parameters which penalize respectively the estimation for the mean and the dispersion function. The choice of the parameters λ_μ and λ_γ is discussed in Section 3.2.

Maximization of (3.1) is done via an iterative two-step procedure. At each iteration i we have the following steps:

STEP (a): for a given $\gamma^{(i-1)}(x)$, choose an optimal value of λ_μ, obtain the maximizer of (3.1) with respect to $\boldsymbol{\alpha}_\mu$ denoted by $\hat{\boldsymbol{\alpha}}_\mu^{(i)}$, and get an estimate $\hat{\mu}^{(i)}(x)$ for $\mu(x)$.

STEP (b): for the given $\hat{\mu}^{(i)}(x)$, choose an optimal value of λ_γ, obtain the maximizer of (3.1) with respect to $\boldsymbol{\alpha}_\gamma$, denoted by $\hat{\boldsymbol{\alpha}}_\gamma^{(i)}$ and, get an estimate $\hat{\gamma}^{(i)}(x)$ for $\gamma(x)$.

The iteration stops when, for a chosen small $0 < \epsilon < 1$, we have that

$$1 - \epsilon < \max_{1 \le j \le n} \left(\left| \frac{\boldsymbol{B}_{\mu,j}^T \boldsymbol{\alpha}_\mu^{(i)}}{\boldsymbol{B}_{\mu,j}^T \boldsymbol{\alpha}_\mu^{(i-1)}} \right|, \left| \frac{\boldsymbol{B}_{\gamma,j}^T \boldsymbol{\alpha}_\gamma^{(i)}}{\boldsymbol{B}_{\gamma,j}^T \boldsymbol{\alpha}_\gamma^{(i-1)}} \right| \right) < 1 + \epsilon \ ,$$

where $\boldsymbol{B}_{\mu,j}^T \boldsymbol{\alpha}_\mu^{(i)}$ (respectively $\boldsymbol{B}_{\gamma,j}^T \boldsymbol{\alpha}_\gamma^{(i)}$) is the estimator for $\mu(x_j)$ (respectively $\gamma(x_j)$) at the i-th iteration.

We now provide more details on the procedure.

- STEP 0: initial values

Some appropriate initial values $\boldsymbol{\alpha_\mu}^{(0)}$ and $\boldsymbol{\alpha_\gamma}^{(0)}$ are needed to start the iterative procedure. In our implementation we have taken constant initial values $\hat{\mu}^{(0)} = n^{-1} \sum_{i=1}^n y_i = \overline{y}$ and $\hat{\gamma}^{(0)} = \phi_n$ where ϕ_n is either the known value for ϕ in the considered exponential model (see Table 2.1 on page 30 of McCullagh and Nelder (1989)) or an estimator for it. For example, for the normal model $\phi = \sigma^2$, and in this case $\hat{\gamma}^{(0)} = (n-1)^{-1} \sum_{i=1}^n (y_i - \overline{y})^2$,

the empirical variance. For a Poisson model $\phi = 1$ and for a binomial model $\phi = 1/N$, where N is the number of trials.

With these initial values the iterative procedure is started and alternates between steps (a) and (b). Within a given iteration i two sub-steps are distinguished: the choice of an optimal smoothing parameter and maximization of (3.1). Details on the optimal choice of λ_μ and λ_γ are in Section 3.2, and more explanation on the maximization of (3.1) with respect to $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ is provided here.

- STEP (a): estimation of $\boldsymbol{\alpha}_\mu$.

To find the maximum likelihood estimator for $\boldsymbol{\alpha}_\mu$ we rewrite (3.1) as a function of $\boldsymbol{\alpha}_\mu$ for a given $\gamma(x)$:

$$l(\boldsymbol{\alpha}_\mu; \boldsymbol{x}, \boldsymbol{y}, \lambda_\mu) = \sum_{i=1}^{n} \left\{ \frac{1}{\phi\, \gamma(x_i)} (y_i\ \boldsymbol{B}_{\mu,i}^T \boldsymbol{\alpha}_\mu - b(\boldsymbol{B}_{\mu,i}^T \boldsymbol{\alpha}_\mu)) \right\} - \frac{1}{2} \lambda_\mu \boldsymbol{\alpha}_\mu^T \boldsymbol{D}_m^T \boldsymbol{D}_m \boldsymbol{\alpha}_\mu. \qquad (3.2)$$

As in the generalized linear models setting of Eilers and Marx (1996) this system cannot be solved analytically, and an iterative Fisher scoring method is needed for maximizing (3.2), or equivalently for solving

$$\boldsymbol{u} = \left(\frac{1}{\phi} \sum_{i=1}^{n} \frac{1}{\gamma(x_i)} \boldsymbol{B}_{\mu,i} (y_i - b'(\boldsymbol{B}_{\mu,i}^T \boldsymbol{\alpha}_\mu)) \right) - \lambda_\mu \boldsymbol{D}_m^T \boldsymbol{D}_m \boldsymbol{\alpha}_\mu = \boldsymbol{0}\ ,$$

where $\boldsymbol{0}$ is the null vector of dimension $K_\mu \times 1$.

The Fisher information matrix $\boldsymbol{F}_\mu$ resulting from (3.2) is defined as

$$\boldsymbol{F}_\mu = \boldsymbol{B}_\mu^T \boldsymbol{W}_\mu \boldsymbol{B}_\mu + \lambda_\mu \boldsymbol{D}_m^T \boldsymbol{D}_m\ ,$$

where $\boldsymbol{W}_\mu$ is the diagonal matrix with elements $\{\phi\gamma(x_i)\}^{-1} b''(\boldsymbol{B}_{\mu,i}^T \boldsymbol{\alpha}_\mu)$. The updating rule for $\boldsymbol{\alpha}_\mu$ is then $\boldsymbol{\alpha}_\mu = \tilde{\boldsymbol{\alpha}}_\mu + \tilde{\boldsymbol{F}}_\mu{}^{-1} \tilde{\boldsymbol{u}}$, where $\tilde{\boldsymbol{\alpha}}_\mu$ is the current approximation for $\boldsymbol{\alpha}_\mu$, and similarly for $\tilde{\boldsymbol{F}}_\mu$ and $\tilde{\boldsymbol{u}}$ (the updated score vector).

Given the current estimate $\tilde{\boldsymbol{\alpha}}_\mu$, we obtain the update

$$\boldsymbol{\alpha}_\mu = (\boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \lambda_\mu \boldsymbol{D}_m^T \boldsymbol{D}_m)^{-1} \boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \tilde{\boldsymbol{z}},$$

with $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \ldots, \tilde{z}_n)^T$ the vector of working variables, whose elements are

$$\tilde{z}_i = \left(\boldsymbol{B}_{\mu,i}^T \tilde{\boldsymbol{\alpha}}_\mu + (y_i - b'(\boldsymbol{B}_{\mu,i}^T \tilde{\boldsymbol{\alpha}}_\mu)) \frac{1}{b''(\boldsymbol{B}_{\mu,i}^T \tilde{\boldsymbol{\alpha}}_\mu)} \right)$$

and the diagonal matrix of weights $\tilde{\boldsymbol{W}}_\mu$ with elements

$$\tilde{w}_{\mu,ii} = \frac{1}{\phi\,\gamma(x_i)} b''(\boldsymbol{B}_{\mu,i}^T \tilde{\boldsymbol{\alpha}}_\mu).$$

This differs from the weight matrix that is used in Fisher scoring methods for the usual estimation of the mean function (see e.g. Eilers and Marx (1996)) in having $\gamma(x_i)$ in the denominator. The weight of each observation is thus defined through both the mean and the dispersion value.

Once convergence is reached, for a given λ_μ, we obtain the vector of estimated coefficients $\hat{\boldsymbol{\alpha}}_\mu(\lambda_\mu)$, and $(\hat{\mu}_{\lambda_\mu}(x_1), \ldots, \hat{\mu}_{\lambda_\mu}(x_n))^T = \boldsymbol{H}_\mu(\lambda_\mu)\hat{\boldsymbol{z}}$, with $\boldsymbol{H}_\mu(\lambda_\mu) = \boldsymbol{B}_\mu(\boldsymbol{B}_\mu^T \hat{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \lambda_\mu \boldsymbol{D}_m^T \boldsymbol{D}_m)^{-1} \boldsymbol{B}_\mu^T \hat{\boldsymbol{W}}_\mu$. For more details see Eilers and Marx (1996).

- Step (b): estimation of $\boldsymbol{\alpha}_\gamma$.

Similar to step (a), the maximum likelihood estimate for $\boldsymbol{\alpha}_\gamma$ is found by rewriting (3.1), for a given $\theta(x)$, as a function of $\boldsymbol{\alpha}_\gamma$:

$$l(\boldsymbol{\alpha}_\gamma; \boldsymbol{x}, \boldsymbol{y}, \lambda_\gamma) = -\frac{1}{2} \sum_{i=1}^{n} \left\{ \log h(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma) + \frac{1}{h(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)} d(y_i, \theta(x_i)) \right\} - \frac{1}{2} \lambda_\gamma \boldsymbol{\alpha}_\gamma^T \boldsymbol{D}_\ell^T \boldsymbol{D}_\ell \boldsymbol{\alpha}_\gamma. \tag{3.3}$$

To maximize (3.3) with respect to $\boldsymbol{\alpha}_\gamma$ we need to solve the following system of K_γ equations:

$$\boldsymbol{v} = -\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{B}_{\gamma,i} \left\{ \frac{h'(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)}{h(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)} - \frac{h'(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)}{h^2(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)} d(y_i, \theta(x_i)) \right\} - \lambda_\gamma \boldsymbol{D}_\ell^T \boldsymbol{D}_\ell \boldsymbol{\alpha}_\gamma = \boldsymbol{0},$$

which needs to be solved iteratively through Fisher scoring.

The Fisher information matrix for (3.3) is a matrix $\boldsymbol{F}_\gamma$

$$\boldsymbol{F}_\gamma = \boldsymbol{B}_\gamma^T \boldsymbol{W}_\gamma \boldsymbol{B}_\gamma + \lambda_\gamma \boldsymbol{D}_\ell^T \boldsymbol{D}_\ell \,,$$

where $\boldsymbol{W}_\gamma$ is the diagonal matrix with elements $2^{-1}\,(h'(\boldsymbol{B}_{\gamma,i}^T \boldsymbol{\alpha}_\gamma)/h(\boldsymbol{B}_{\gamma,i} \boldsymbol{\alpha}_\gamma))^2$.

After some algebra we find that the updating rule $\boldsymbol{\alpha}_\gamma = \tilde{\boldsymbol{\alpha}}_\gamma + \tilde{\boldsymbol{F}}_\gamma^{-1} \tilde{\boldsymbol{v}}$ results into

$$\boldsymbol{\alpha}_\gamma = \left(\boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \lambda_\gamma \boldsymbol{D}_\ell^T \boldsymbol{D}_\ell \right)^{-1} \boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \tilde{\boldsymbol{q}} \,,$$

with $\tilde{\boldsymbol{\alpha}}_\gamma$ the current estimate, $\tilde{\boldsymbol{W}}_\gamma$ the diagonal matrix of current weights

$$\tilde{w}_{\gamma,ii} = \frac{1}{2} \left(\frac{h'(\boldsymbol{B}_{\gamma,i}^T \tilde{\boldsymbol{\alpha}}_\gamma)}{h(\boldsymbol{B}_{\gamma,i} \tilde{\boldsymbol{\alpha}}_\gamma)} \right)^2 ,$$

and $\tilde{\boldsymbol{q}} = (\tilde{q}_1, \ldots, \tilde{q}_n)^T$ the current working variable vector with components

$$\tilde{q}_i = \left(\boldsymbol{B}_{\gamma,i}^T \tilde{\boldsymbol{\alpha}}_\gamma + \big(d(y_i, \theta(x_i)) - h(\boldsymbol{B}_{\gamma,i}^T \tilde{\boldsymbol{\alpha}}_\gamma) \big) \frac{1}{h'(\boldsymbol{B}_{\gamma,i}^T \tilde{\boldsymbol{\alpha}}_\gamma)} \right).$$

Once convergence is reached we obtain, similar as for the mean function, a final parameter estimate $\hat{\boldsymbol{\alpha}}_\gamma(\lambda_\gamma)$, and we build a hat matrix $\boldsymbol{H}_\gamma(\lambda_\gamma) = \boldsymbol{B}_\gamma(\boldsymbol{B}_\gamma^T \hat{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \lambda_\gamma \boldsymbol{D}_\ell^T \boldsymbol{D}_\ell)^{-1} \boldsymbol{B}_\gamma^T \hat{\boldsymbol{W}}_\gamma$ such that $(\hat{\gamma}_{\lambda_\gamma}(x_1), \ldots, \hat{\gamma}_{\lambda_\gamma}(x_n))^T = \boldsymbol{H}_\gamma(\lambda_\gamma)\hat{\boldsymbol{q}}$.

3.2 The choice of the smoothing parameters

The two smoothing parameters λ_μ and λ_γ are selected (within the iteration steps) via generalized cross validation. See Wahba (1990) for a standard reference. A value for λ_μ is obtained by minimizing

$$\mathrm{GCV}(\lambda_\mu) = \sum_{i=1}^{n} \frac{d(y_i, \hat{\theta}_{\lambda_\mu}(x_i))/\gamma(x_i)}{(n - \mathrm{df}(\lambda_\mu))^2}, \qquad \text{with} \quad \mathrm{df}(\lambda_\mu) = \mathrm{tr}(\boldsymbol{H}_\mu(\lambda_\mu)), \tag{3.4}$$

where the numerator is based on the appropriate deviance.

For choosing λ_γ we mimic (3.4), recalling the expression for the log-likelihood in (3.3) and writing down the appropriate deviance. Taking $\mathrm{df}(\lambda_\gamma) = \mathrm{tr}(\boldsymbol{H}_\gamma(\lambda_\gamma))$ which can be interpreted as the equivalent number of degrees of freedom needed when fitting the dispersion function, we define

$$\mathrm{GCV}(\lambda_\gamma) = \sum_{i=1}^{n} \frac{\log \hat{\gamma}_{\lambda_\gamma}(x_i) + d(y_i, \theta(x_i))/\hat{\gamma}_{\lambda_\gamma}(x_i) - \log d(y_i, \theta(x_i)) - 1}{(n - \mathrm{df}(\lambda_\gamma))^2}. \tag{3.5}$$

We then select the value of λ_γ that minimizes $\mathrm{GCV}(\lambda_\gamma)$.

Practically, though, (3.4) and (3.5) cannot be minimized analytically, and we need to use a recursive refinement method to find the optimal values for λ_μ and λ_γ. For brevity we only discuss the search for λ_μ (the search for λ_γ goes along similar lines).

We specify a starting grid of λ_μ values and obtain, for each value of λ_μ, the different estimates $\hat{\boldsymbol{\alpha}}_\mu(\lambda_\mu)$, and consequently $\hat{\mu}_{\lambda_\mu}$ and $\mathrm{df}(\lambda_\mu)$. We then find the value $\lambda_\mu^{(1)}$ that minimizes $\mathrm{GCV}(\lambda_\mu)$ over the grid of λ_μ-values. Subsequently we create a new grid, centered around $\lambda_\mu^{(1)}$, with a smaller range than the previous one, and find the next value for which the GCV value is minimal. We keep on refining more and more the grid around the chosen value until we find a $\lambda_\mu^{(k)}$ such that: $1 - \epsilon_\lambda < (\lambda_\mu^{(k)}/\lambda_\mu^{(k-1)}) < 1 + \epsilon_\lambda$, with ϵ_λ a chosen small value. In the numerical studies in Sections 4 and 5 we took $\epsilon_\lambda = 0.05$.

4 Simulation studies

The estimation procedure proposed in Section 3 provides a flexible tool for analyzing data from distributions belonging to the double exponential family, allowing for overdispersion and/or underdispersion to vary with a covariate. In this section we investigate the performance of the method via a simulation study, involving data from a (double) normal distribution, a double Poisson distribution and a double binomial distribution. In Section 4.4 we provide comparisons with other existing methods.

Throughout this and the next section we use the canonical link function for the estimation of the mean and for the dispersion function we take $\gamma(x) = h(\boldsymbol{B}_\gamma^T(x)\boldsymbol{\alpha_\gamma}) = \exp(\boldsymbol{B}_\gamma^T(x)\boldsymbol{\alpha_\gamma})$. We take $m = \ell = 2$ for the order of the difference operators in the penalty terms. The number of simulations is $M = 1000$. The number and locations of knots are fixed throughout all simulations for a same model.

To evaluate the quality of the estimated functions, we calculate the approximate integrated squared error (AISE)

$$\text{AISE}^{(s)} = \frac{\sum_{x_{\text{grid}}} \left(\hat{f}^{(s)}(x_{\text{grid}}) - f_{\text{true}}(x_{\text{grid}})\right)^2}{\sum_{x_{\text{grid}}} \left(f_{\text{true}}(x_{\text{grid}})\right)^2}, \quad \text{for } s = 1, \ldots, M,$$

for each of the M simulations (indexed by s). Herein x_{grid} is an appropriate grid of values, $f_{\text{true}}(\cdot)$ and $\hat{f}(\cdot)$ are, respectively, the true and the estimated function (either the mean or the dispersion function).

All computing has been done using the R-programming environment. Computer source codes for the discussed statistical procedure are made available at the web page http://wis.kuleuven.be/stat/codes.html

4.1 Normal data

In this model $\mu(x) = 8.2 + 45\sin(3x)(\cos(7x))^2$ and $\gamma(x) = 12 + 8\sin(-5x + 0.1)(\cos(8x + 0.1))^2 + 2\log(4.2x + 0.1)$. See Figure 4.1. We apply the estimation method using B-splines of degrees $p_\mu = p_\gamma = 3$, with a set of equally-spaced knots of dimensions $k_\mu = 40$ and $k_\gamma = 34$. From 1000 simulated samples, we sort the 1000 AISE-values, and present in Figure 4.1 the estimated mean and dispersion (variance) function corresponding to the 5th, the 50th and the 95th quantile of the AISE-values. From Figure 4.1 it can be seen that the estimates catch quite well the shape of the true curves. The estimation of the variance function is more difficult than that of the mean function, as expected.

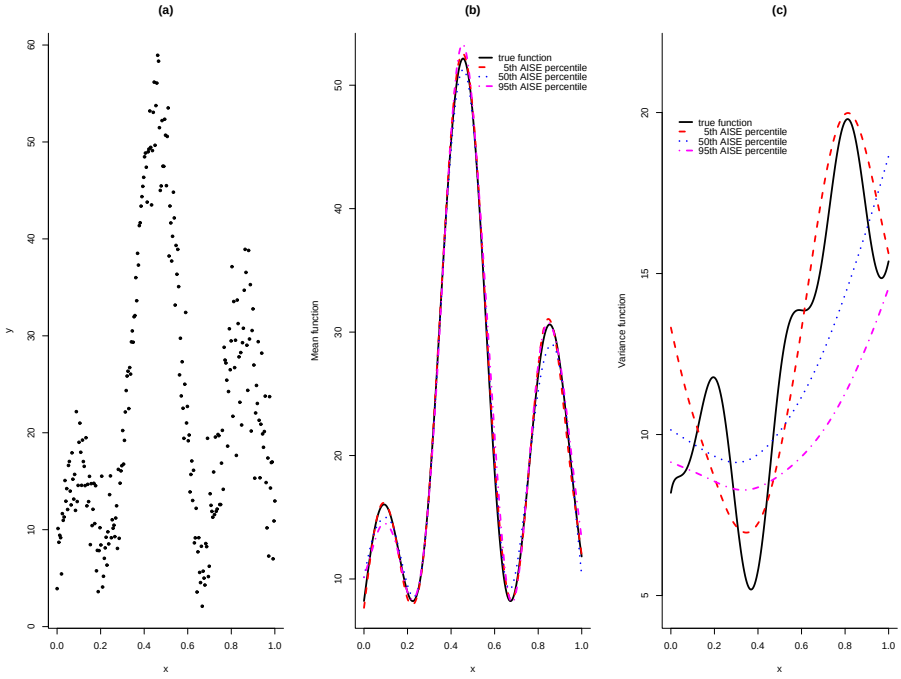


Figure 4.1: *Normal model with $n = 250$. (a) a simulated dataset; (b) the true mean function and (c) the true variance function, with representative estimates.*

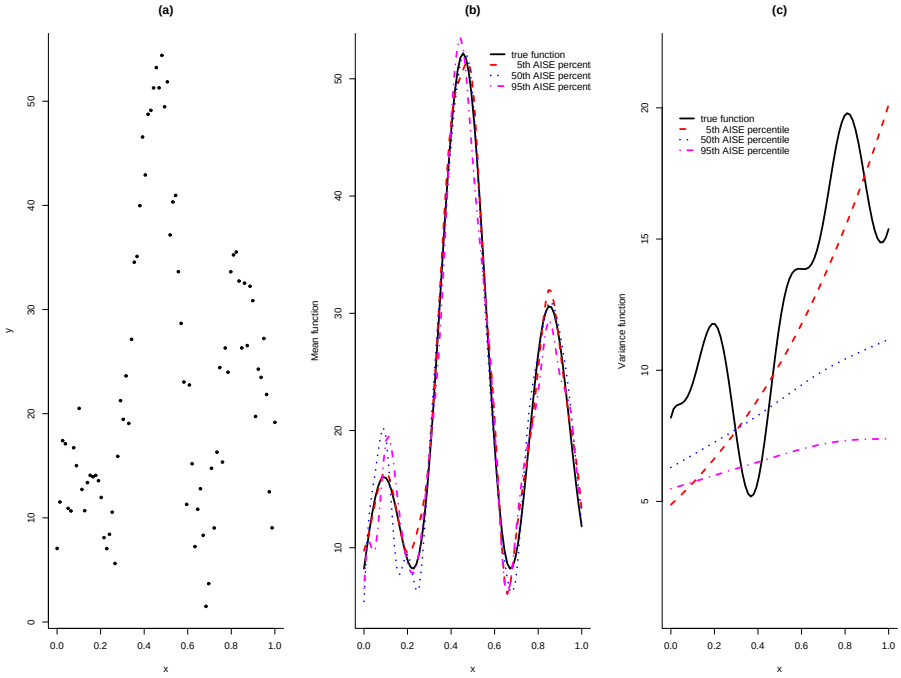


Figure 4.2: *Normal model with $n = 80$. (a) a simulated dataset; (b) the true mean function and (c) the true variance function, with representative estimates.*

Figure 4.2 depicts similar results as Figure 4.1, but now based on samples of size $n = 80$. Here we took $p_\mu = p_\gamma = 3$ with $k_\mu = 20$ and $k_\gamma = 18$. We can see that, even for

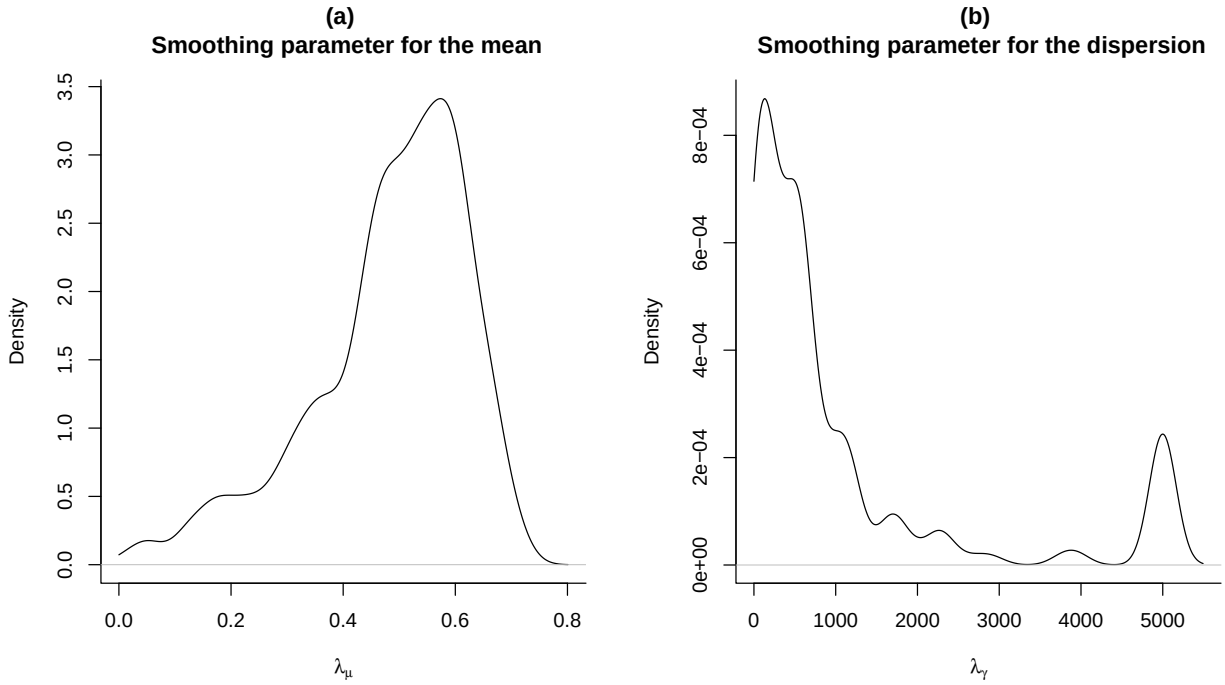


Figure 4.3: *Normal model with* $n = 250$. *Estimated density for the GCV-selected* λ_μ *and* λ_γ *parameters.*

this smaller sample size, the estimation of the mean function is still very accurate. The estimation of the variance function however is far less precise. Although, in general, the estimate catches the trend in the function, it often misses the bumps.

In Figure 4.3 we present the density of the selected smoothing values λ_μ and λ_γ, obtained via the GCV-criteria as described in Section 3.2, for the 1000 samples of size $n = 250$. Note that the selected λ_γ-values are far larger in magnitude and more spread out than the selected λ_μ-values.

Of interest is also to look for constructing approximate confidence intervals. A possibility is to rely on the Fisher information matrices derived in Section 3, and a large sample approximation by a normal distribution with mean the estimated target function and variance the generalization of the variance of a generalized linear model (as in formula (7) of Marx and Eilers (1998)). Confidence intervals for the components $\eta(\cdot)$ and $\xi(\cdot)$ are presented in Figure 4.4 as short-dashed curves. Plotted are also the true functions (as solid curves) as well as the estimated curve corresponding to a median AISE performance (dotted curve). From Figure 4.4 (b) it is again seen that estimation of the dispersion function is a far more difficult problem than estimation of the mean function.

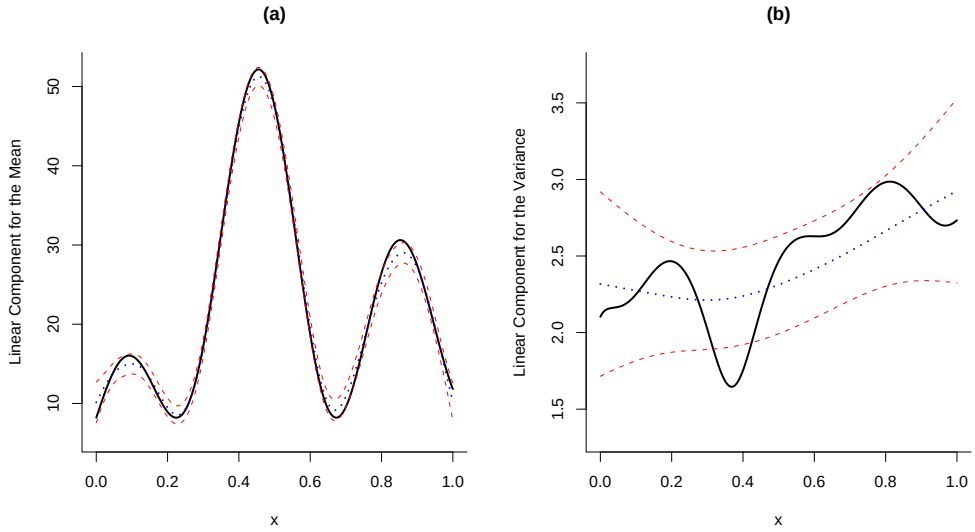


Figure 4.4: *Normal model with $n = 250$: the estimate of the linear components $\eta(\cdot)$ and $\xi(\cdot)$ corresponding to the median AISE values (dotted curve) for the mean (a) and the variance (b) function with 95% confidence intervals (short-dashed curves).*

4.2 Poisson data

We now simulate from two different double Poisson models. For both models we take the same dispersion function $\gamma(x) = 1.7 - 0.3x^2 + 0.4\sin(6x)$ but we consider two different mean functions. The mean function for the setting of Figure 4.5 equals $\mu(x) = 26 + 40x + 48(x - 0.6)^2 + 16\sin(6x)$ and is 25 times the mean function in the setting of Figure 4.6. The true mean, dispersion and variance functions for both settings are depicted as solid curves in panels (b)—(d) of Figures 4.5 and 4.6. Typical simulated datasets are in panels (a) of Figures 4.5 and 4.6. We perform 1000 simulations of sample size $n = 210$. For the estimation procedure we use B-spline basis functions of degree $p_\mu = p_\gamma = 3$ and equally-spaced knots with $k_\mu = 45$ and $k_\gamma = 30$. For both models, the method gives satisfactory results: the general shape of the functions is well described by their estimates.

In Figure 4.7 approximate confidence intervals for the Poisson model with high mean function are presented. Note again the difference in width between the confidence intervals for the function $\eta(\cdot)$ on the one hand and the function $\xi(\cdot)$ on the other hand.

4.3 Binomial data

We now consider two binomial models. In presenting (and in analyzing) the data we use the proportion, rather than the actual number of successes over a certain number of trials in each experiment. The mean function is $\mu(x) = \exp(-2.5 + 7.5x^2 + 3.3x^2\sin(6x)/(1 +$

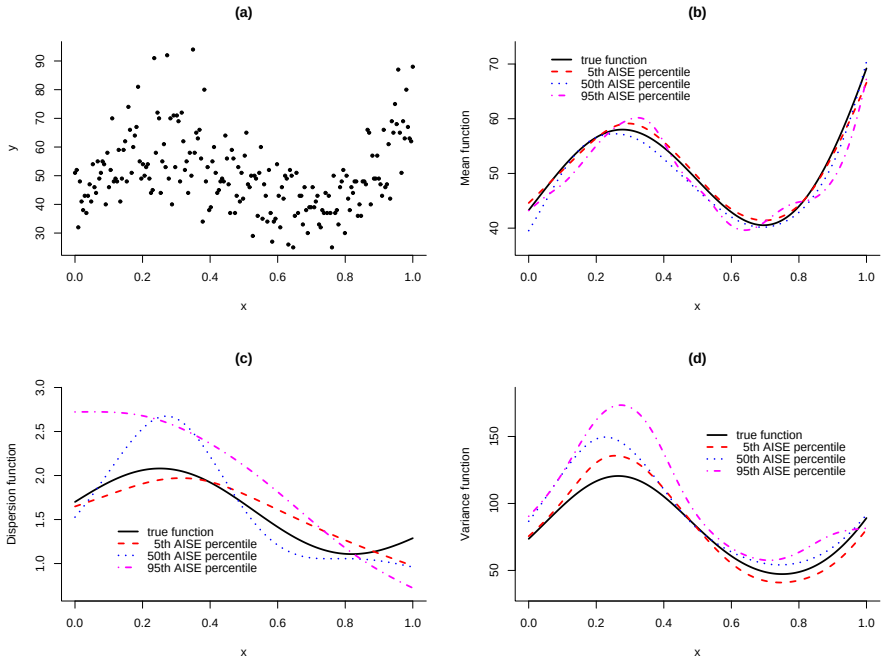


Figure 4.5: *Poisson model with $n = 210$. (a) a simulated dataset, (b) the true (high) mean function, (c) the true dispersion function and (d) the true variance function, with representative estimates.*

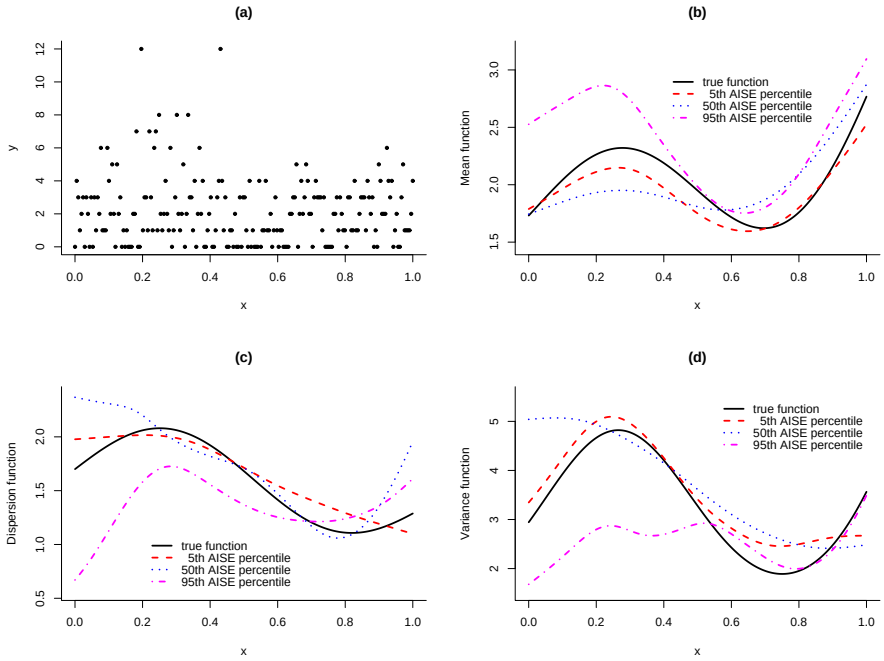


Figure 4.6: *Poisson model with $n = 210$. (a) a simulated dataset, (b) the true (low) mean function, (c) the true dispersion function and (d) the true variance function, with representative estimates.*

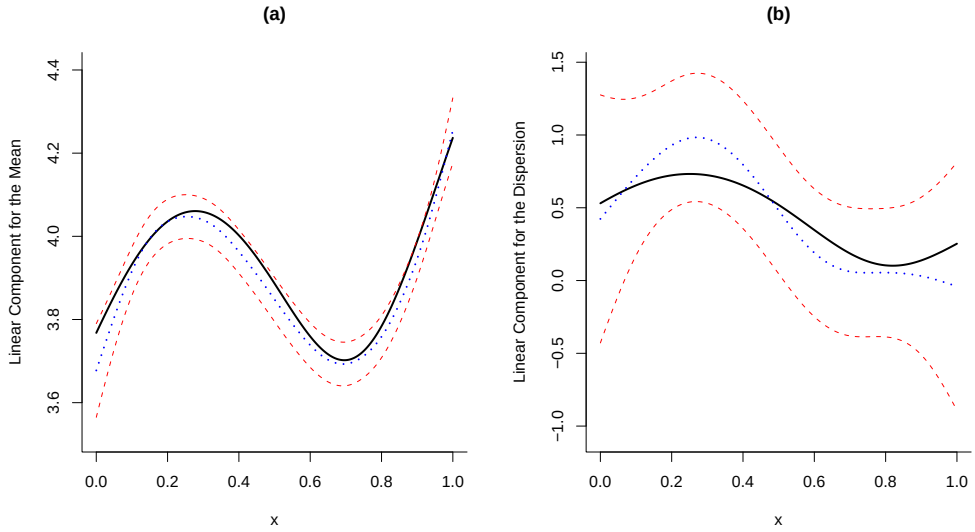


Figure 4.7: *High mean Poisson model with* $n = 210$*: the estimate of the components* $\eta(\cdot)$ *and* $\xi(\cdot)$ *corresponding to the median AISE values (dotted curve) for the mean (a) and the dispersion (b) function with 95% confidence intervals (short-dashed curves).*

$\exp(-2.5 + 7.5x^2 + 3.3x^2 \sin(6x))$ and the dispersion function equals $\gamma(x) = \exp(0.8 - 0.8\cos(4x+0.2)\sin(6x))$. The underlying mean, dispersion and variance function of a first model are depicted as solid curves in panels (b)—(d) in Figure 4.8. In this first model the constant number of trials is equal to 220. We draw 1000 samples of size $n = 200$ from this model. A typical simulated dataset is depicted in Figure 4.8 (a). In the estimation procedure we take B-spline basis functions with $p_\mu = p_\gamma = 3$ and $k_\mu = 45$ and $k_\gamma = 38$ knots. The estimated curves corresponding to the 5th, the 50th and the 95th AISE-quantiles are presented in Figure 4.8. Again estimation of the mean is more precise than estimation of the dispersion function, but nevertheless the results are quite satisfactory. For a second binomial case we reduce the number of trials to 20. The results, obtained with the same B-spline bases functions and the same choices of knots, are still encouraging although less precise (see Figure 4.9).

Finally, confidence intervals for the function $\eta(\cdot)$ and $\xi(\cdot)$ are provided in Figure 4.10. The conclusions are similar as for the previous models.

4.4 Comparison with other methods

The approach of tackling estimation of mean and dispersion via the double exponential family and P-splines regression is quite appealing, since it allows for flexible modeling and can handle overdispersion as well as underdispersion. In this section we compare the

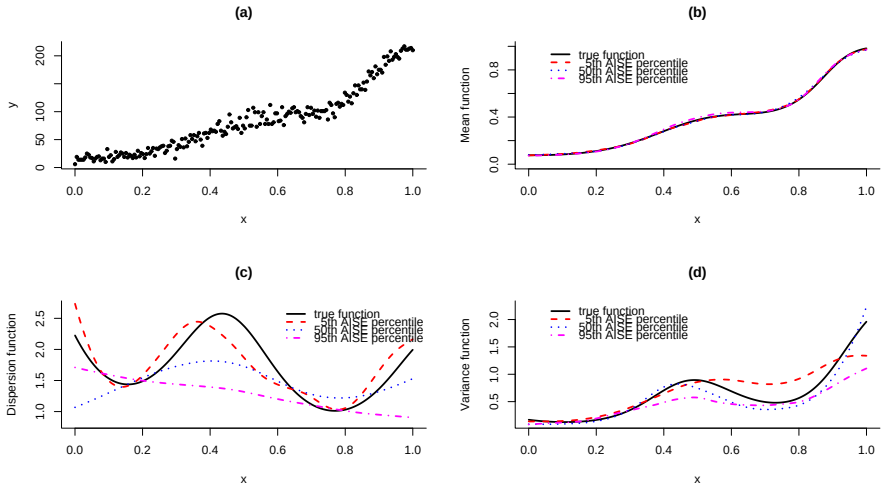


Figure 4.8: *Binomial model with* $n = 200$ *and number of trials is 220. (a) a simulated dataset, (b) the true mean function, (c) the true dispersion function and (d) the true variance function, with representative estimates.*

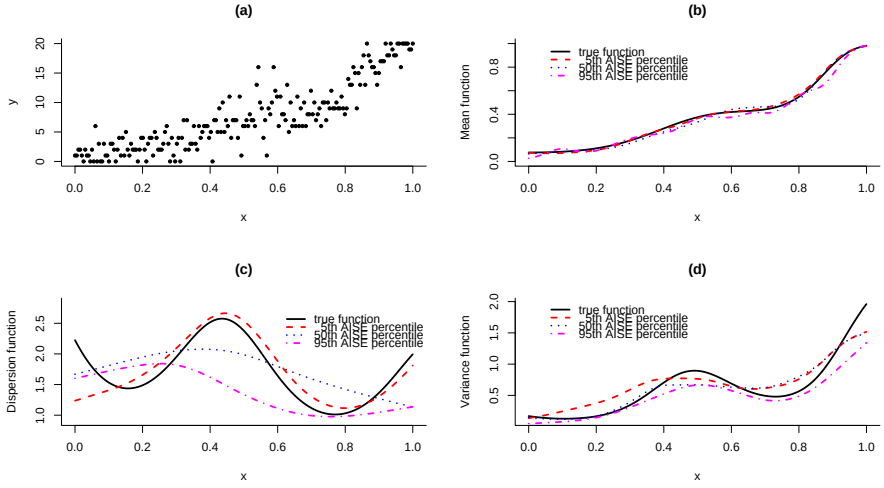


Figure 4.9: *Binomial model with* $n = 200$ *and number of trials is 20. (a) a simulated dataset, (b) the true mean function, (c) the true dispersion function and (d) the true variance function, with representative estimates.*

proposed estimation method with other methods for variance or dispersion estimation available in the literature.

A first question one might ask is what can be gained by allowing for a dispersion that changes with the covariate value x, as opposed to just modeling an additional dispersion parameter.

A second question is how the proposed method compares to other methods for estimating a dispersion function (or a variance function in the normal model). In this section

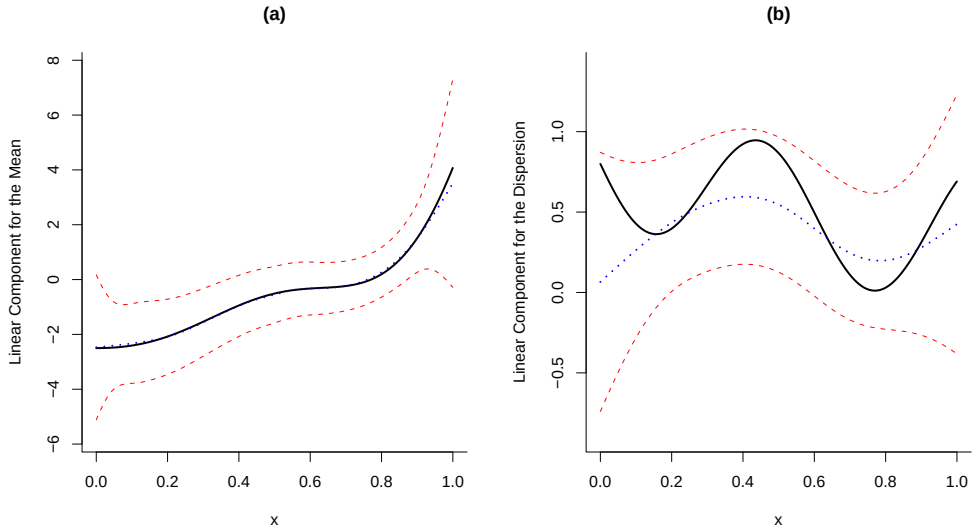


Figure 4.10: *Binomial model with* $n = 200$ *and the number of trials is 220: the estimate of* $\eta(\cdot)$ *and* $\xi(\cdot)$ *corresponding to the median AISE values (dotted curve) for the mean (a) and the dispersion (b) function with 95% confidence intervals (short-dashed curves).*

we compare our approach with the mixed model approach of Rigby and Stasinopoulos (2005). For the normal model, when estimating a variance function, we also compare with a difference type of variance estimation. See for example Hall *et al.* (1990).

In difference-based estimation of the variance function, one needs to choose a difference sequence $\{d_j\}$, which is a sequence of real numbers such that $\sum d_j = 0$ and $\sum d_j^2 = 1$. For a given strictly positive integer m, put $d_j = 0$ for $j > m$, and calculate the sequence of differences of subsequent Y-observations $(v_1, \cdots, v_{n-m})$ where $v_i = \sum_{j=0}^{m} d_j y_{i+j}$ for $i = 1, \cdots n - m$. Variance estimators are then based on the sequence $(v_1, \cdots, v_{n-m})$, with the simplest example its empirical variance. In the comparison we use a P-spline to estimate the variance from the sequence $(v_1, \cdots, v_{n-m})$. This difference-based variance estimation depends on the choice of the difference order m. We present results for the method when fixing the order $m = 3$, but for comparison reasons also include results obtained when choosing the optimal m from the set of values $1, 2, \cdots, 6$. More precisely, for each sample we calculate the difference-based estimator for all values of m, and report on the estimator for which the AISE value if minimal. Note that when doing so, we use the information on the true function when selecting the optimal m for the sample at hand.

Rigby and Stasinopoulos' (2005) GAMLSS framework also allows for estimation of the mean and the variance/dispersion function. The `gamlss` function of the `gamlss` R package (Stasinopoulos and Rigby (2007)) though uses by default smoothing splines with a fixed

number of degrees of freedom. For comparison purpose we also use P-splines in the Rigby and Stasinopoulos (2005) method applying the `pb` function, taking the same number of knots, the same degree and the same order of difference operator in both methods.

When dealing with overdispersed data one possibility is to estimate the one-parameter exponential family dispersion parameter ϕ as a constant via the deviance residuals. If the interest instead is to estimate the dispersion as a function of the covariate, different methods can be used. A common approach is to make some distributional assumptions on the location parameter so that the data are modeled via specific two-parameter distributions which allow the dispersion to be non constant. In particular overdispersed count data are modeled via a negative binomial distribution and overdispersed proportion data via a beta-binomial distribution. These modeling approaches are also used by Rigby and Stasinopoulos (2005). Although the double exponential family and the hierarchical models both allow to model the dispersion, the actual assumed distribution can be quite different. For a full and fair comparison between the different methods, we therefore present results when simulations are done under the two different schemes (double exponential family and either a negative binomial or a beta-binomial) in such a way that the mean and variance function would be the same.

To summarize, we present results on AISE values for the following methods:

`Const.Var` or `Const.Disp`: results obtained from the proposed method but assuming the variance or dispersion to be a constant;

DEF: proposed estimation procedure, starting from a double exponential family;

RS: method of Rigby and Stasinopoulos (2005);

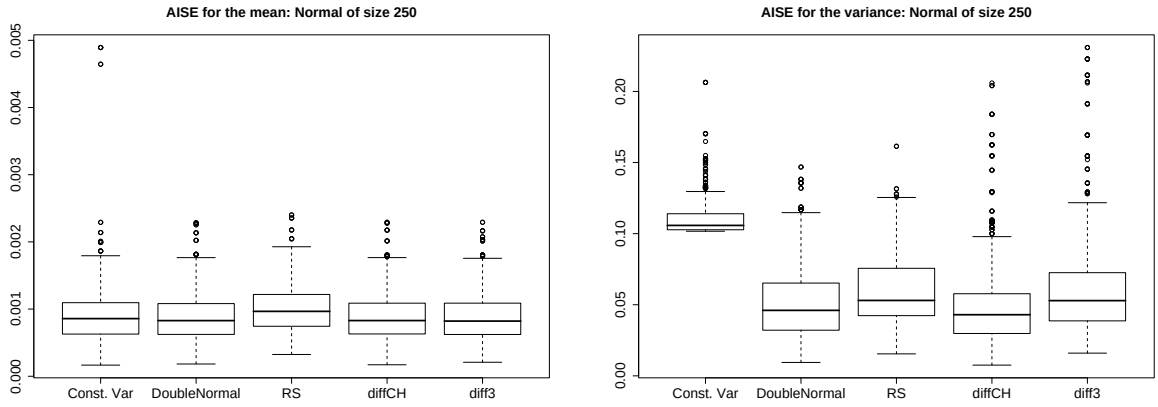
`diffCh`: difference-based estimation of the variance with optimal choice of the difference order m in each simulated sample;

`diff3`: difference-based estimation of the variance with $m = 3$,

when either generating data from an double exponential family (DEF) model or from a negative binomial or a beta-binomial (for the Poisson or binomial case respectively).

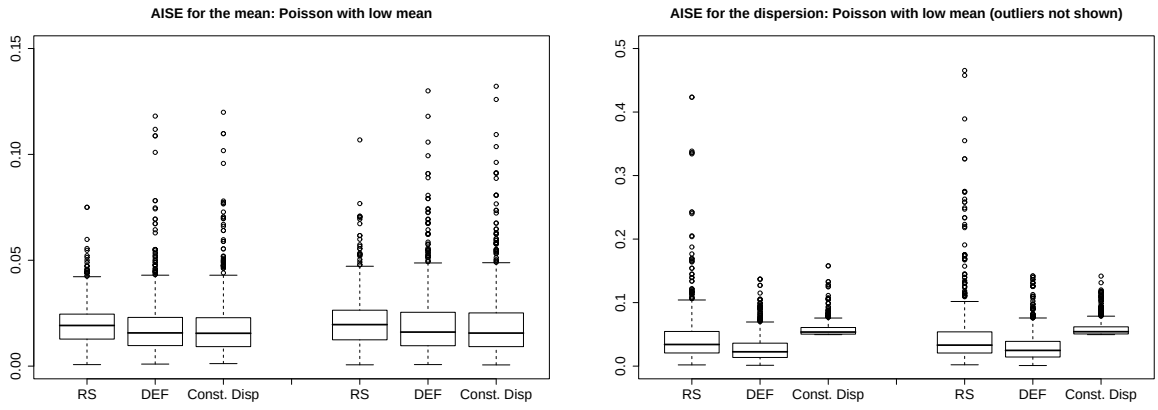
The boxplots reporting on the comparisons of the methods are given in Figures 4.11—4.13. For the normal model, it can be seen from Figure 4.11 (a) that estimation of the mean function is only slightly affected by the method. More differences can be noticed for the variance estimation in Figure 4.11 (b): estimating the variance function as a constant gives very high AISE values. Not surprising is also that the difference-based variance estimation

(`diffCH ) has lower AISE values. The other three estimation`procedures give comparable results, with the double normal method performing somewhat better than the others.



(a) (b)

Figure 4.11: *Normal model with $n = 250$: boxplots of the AISE values for the mean (a) and the dispersion (b) for different variance estimation procedures.*

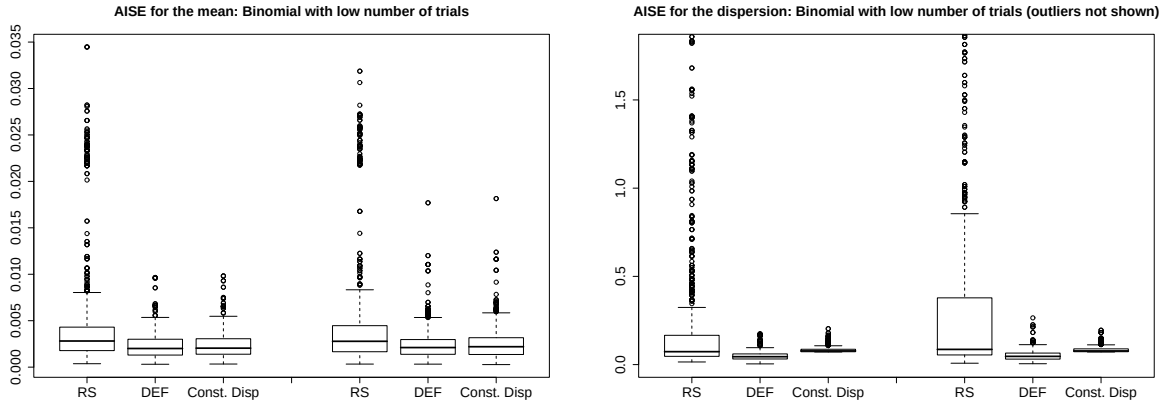


(a) (b)

Figure 4.12: *Low mean Poisson model with $n = 210$: boxplots of the AISE values for the mean (a) and the dispersion (b) for different dispersion estimation procedures. Data are generated either via a DEF (left group of boxplots) or a negative binomial (right group of boxplots).*

In Figure 4.12 results are presented for the Poisson model with low mean. First of all we notice that the performance of the different methods is comparable whether the data are generated via a DEF or a negative binomial. Also, not surprisingly, estimating the dispersion as a constant always leads to higher AISE values for the dispersion, and in the

case of higher mean (results not reported here) estimating the dispersion as a function rather than as a constant, also leads to noticeably better performance in estimation of the mean function. Comparing the hierarchical approach of Rigby and Stasinopoulos (2005) with the proposed DEF approach we see that the performances are quite comparable with a slight dominance of the proposed approach.



(a) (b)

Figure 4.13: *Binomial model with* $n = 200$ *and number of trials is* 20*: boxplots of the AISE values for the mean (a) and the dispersion (b) for different dispersion estimation procedures. Data are generated either via a DEF (left group of boxplots) or a beta-binomial (right group of boxplots).*

As far as proportion data are concerned we can see in Figure 4.13 the effect of the different dispersion estimation methods on data generated either via a DEF or a beta-binomial with low number of trials. Here the DEF approach performs much better than the beta-binomial approach for both simulation schemes. It should be mentioned that in order to have a variance for the data corresponding to the true known function, the true values for the scale parameter of a beta-binomial are very close to the limiting value and vary very little. This for sure makes the estimation more complex and could partially explain the poor performance. At the same time it also sheds some light on how different the two approaches are. In particular values of γ close to 1 correspond to the limiting value of the scale parameter of the beta-binomial distribution and in general the beta-binomial approach would behave badly in cases in which part of the data would be underdispersed. In contrast, the proposed approach can deal with underdispersion, overdispersion and even a combination of these.

5 Data examples

We now illustrate the proposed estimation procedure on several datasets. Again we present results for continuous data, counts and proportions.

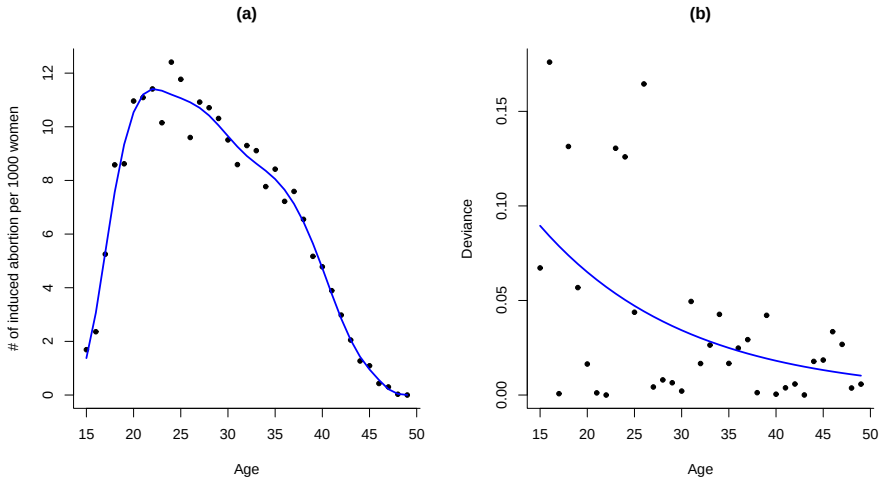


Figure 5.1: *The Italian induced abortions data: (a) mean and (b) dispersion function estimation ($k_\mu = 6$ $k_\gamma = 5$, $p_\mu = 3$ $p_\gamma = 3$).*

5.1 The Italian Induced Abortions data

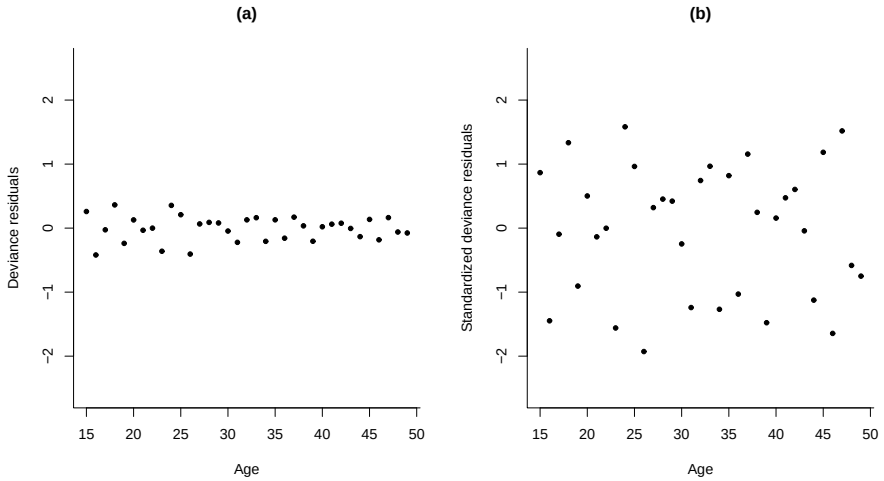


Figure 5.2: *The Italian induced abortions data: (a) unstandardized and (b) standardized deviance residuals.*

The National Institute of Statistics in Italy (ISTAT) collects different information on induced abortion in order to monitor the phenomenon. Among the collected information

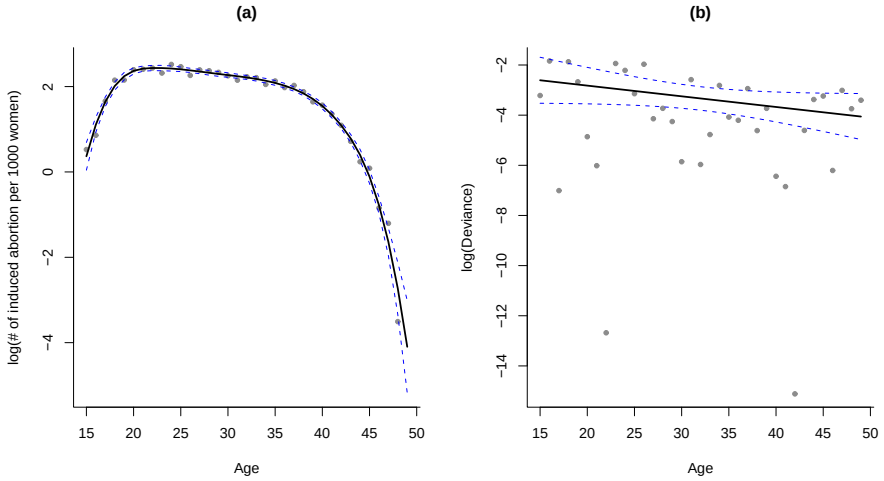


Figure 5.3: *The Italian induced abortions data. The estimated functions (a) $\eta(\cdot)$ and (b) $\xi(\cdot)$ (solid curves) with approximate confidence intervals (dashed curves).*

is the total number of induced abortions on women of a specific age for each Italian region in the year 2006. We present here data for the Veneto region standardized in such a way that we have the counts of induced abortions per 1000 women as a dependent variable. The data, with an estimate for the mean, and the calculated deviance residuals, with an estimate of the dispersion function, are shown in Figure 5.1. When using a Poisson model we expect the variance of the data to be the same as the mean, but here we notice that the variance of the data is always considerably smaller than the mean (underdispersion). From the dispersion function estimate in Figure 5.1 (b) it is seen that there is a quite strong underdispersion over the whole covariate range. In Figure 5.2 we present the unstandardized (considering the dispersion as constant) and the standardized residuals (standardized using the estimated dispersion). Figure 5.2 (b) gives a clear indication that the estimation of the dispersion function performed quite well.

In Figure 5.3 (a) the estimated function $\hat{\eta}(\cdot)$ and its corresponding approximate confidence intervals are given. Figure 5.3 (b) depicts the estimation for $\xi(\cdot)$ with approximate confidence intervals.

5.2 The Fabric data

A classical dataset for overdispersed Poisson data is the fabric dataset, containing 32 observations of the number of faults in rolls of fabric, and the logarithm of the length of each roll. See Hinde (1982). The data, with an estimate for the mean, and the calculated deviance residuals, with an estimate of the dispersion function, are shown in Figure 5.4.

The deviance residuals seem indeed to be dependent on the length of the roll. In Figure 5.5, we plot the unstandardized and standardized residuals. Figure 5.5 (b) shows no or less dependence on the covariate. Note that in this data example, the estimated dispersion function $\hat{\gamma}(x)$ goes from values less than one to values bigger than one, i.e. is crossing over from underdispersion to overdispersion.

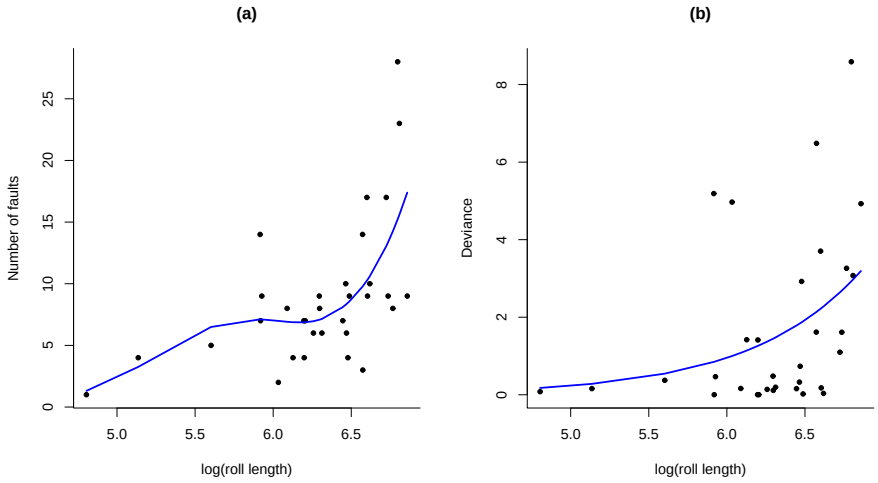


Figure 5.4: *The Fabric data: (a) mean and (b) dispersion function estimation ($k_\mu = 7$ $k_\gamma = 5$, $p_\mu = 3$ $p_\gamma = 2$).*

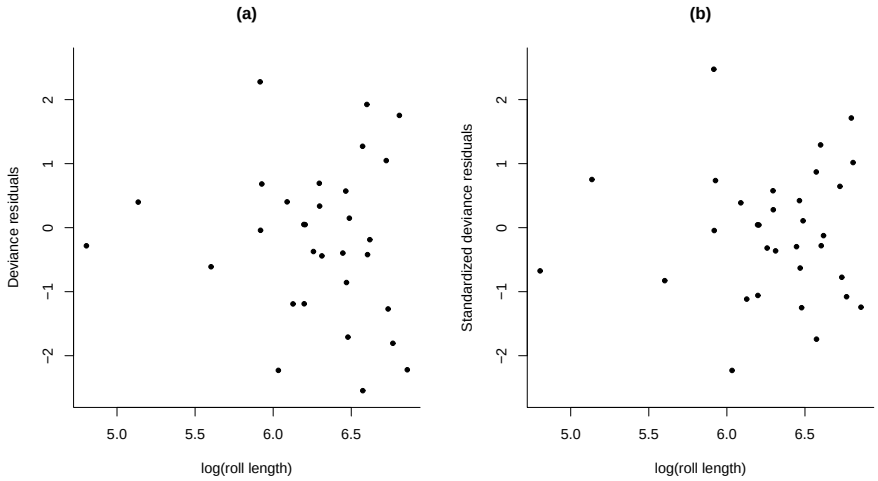


Figure 5.5: *The Fabric data: (a) unstandardized and (b) standardized deviance residuals.*

5.3 The Hospital stay data

The Hospital stay data concern a study at the Hospital del Mar in Barcelona during the years 1988 and 1990. See Alonso *et al.* (1996). The data set contains, among

other measurements, the number of inappropriate days spent out of the whole stay in the hospital in the year 1988 (750 patients) and 1990 (633 patients). For this illustration we focus our attention on the data concerning year 1988, and we are interested in studying the relationship between the age of the patient and the proportion of inappropriate days out of all days spent in the hospital. We model the data as a double binomial family. The estimated mean and dispersion function are shown in Figure 5.6. Figure 5.7 (a) indicates that the variance of the data increases with the age of the patients, while the standardized residuals Figure 5.7 (b) no longer shows this dependence.

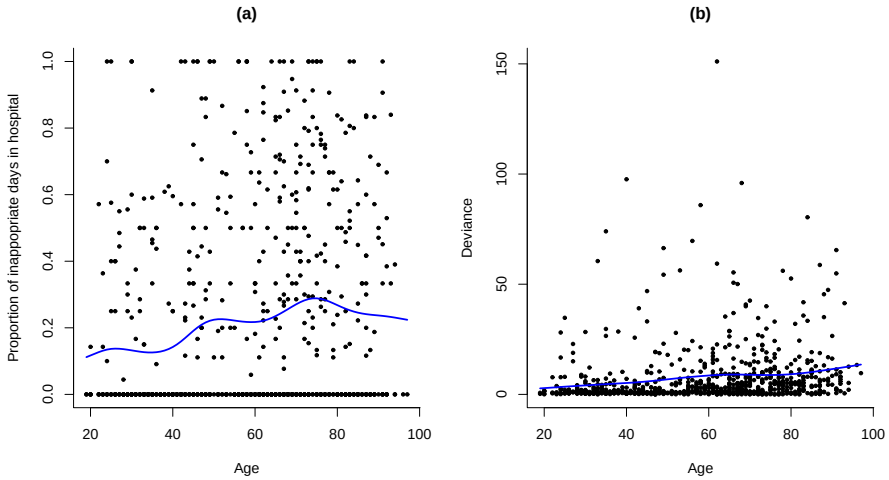


Figure 5.6: *The Hospital stay data: (a) mean and (b) dispersion function estimation ($k_\mu = 52$ $k_\gamma = 44$, $p_\mu = 3$ $p_\gamma = 2$).*

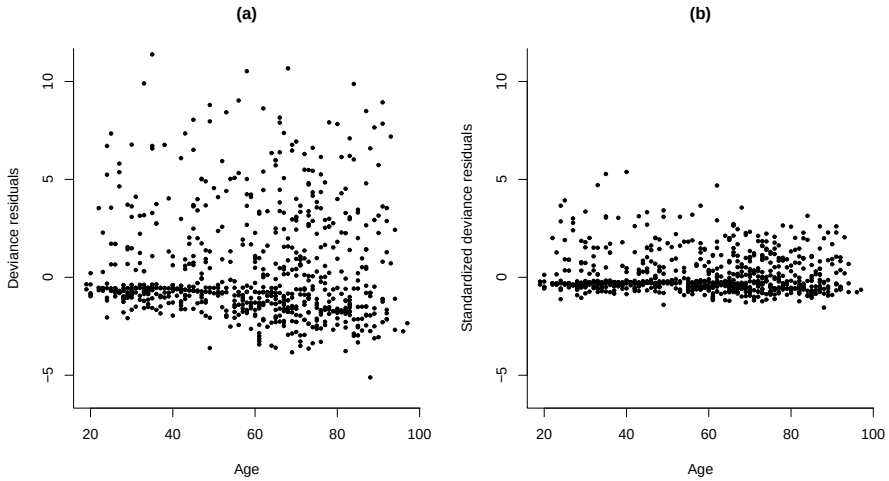


Figure 5.7: *The Hospital stay data: (a) unstandardized and (b) standardized deviance residuals.*

5.4 The low iron rat data

A known dataset concerning overdispersed binomial data is the low iron rat data described by Moore and Tsiatis (1991). An experiment is conducted to see whether the level of hemoglobin in the blood has an effect on the number of dead fetus found in the litter. 58 rats are given an iron supplement, at different dose levels; after 3 weeks they are sacrificed and the proportion of dead fetus on the total size of the litter is recorded. The data, with an estimate for the mean and the dispersion functions are shown in Figure 5.8. The decreasing shape of the deviance seems to be well captured by the estimated function. The plots of the unstandardized and the standardized residuals, in Figure 5.9, convince us of the necessity of modeling the dispersion function as a function of the covariate.

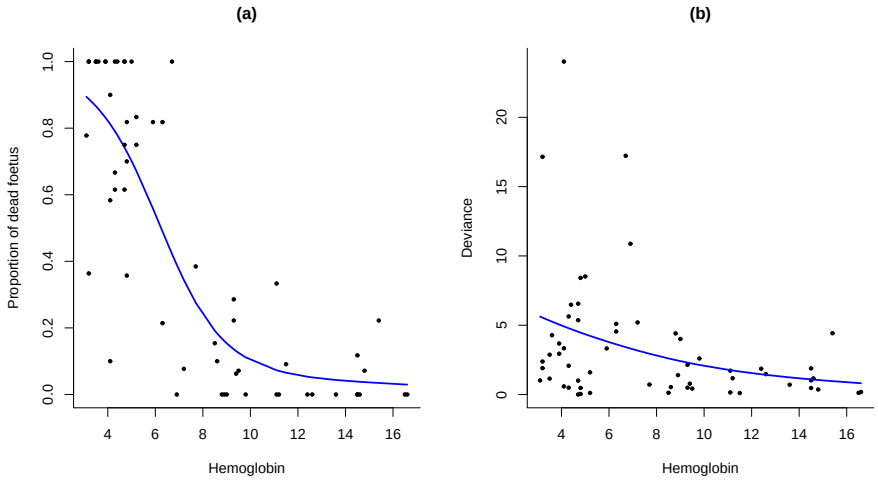


Figure 5.8: *The low iron rat data: (a) mean and (b) dispersion function estimation* ($k_\mu = k_\gamma = 11$, $p_\mu = p_\gamma = 2$).

These data have been also analyzed by Aerts and Claeskens (1997) via a local beta-binomial model and by Nott (2006), who used a DEF approach with Bayesian semiparametric fitting. The estimated mean and dispersion functions seen in Figure 5.8 are very similar to the ones found by the other authors.

The estimated curves $\eta(\cdot)$ and $\xi(\cdot)$ together with their approximate confidence intervals are depicted in respectively Figure 5.10 (a) and Figure 5.10 (b). Note that the pointwise confidence intervals for $\eta(\cdot)$ are wider at the left end of the support (since there the variability in the data is larger).

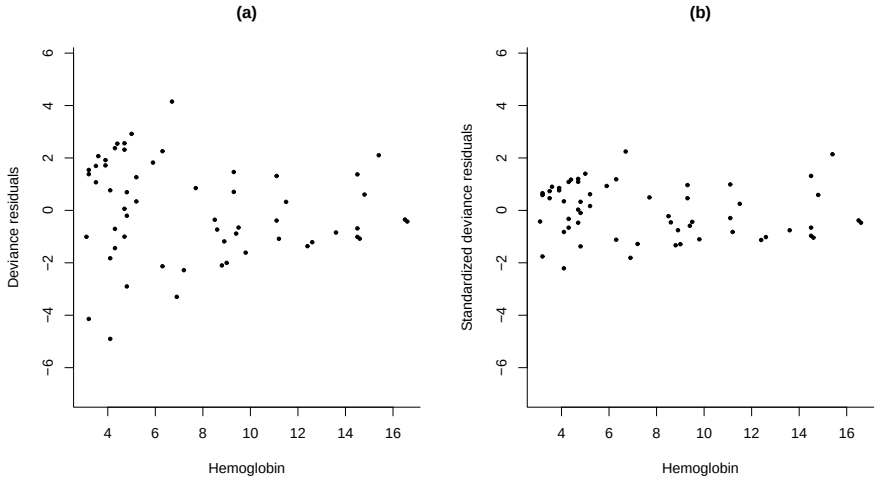


Figure 5.9: *The low iron rat data: (a) unstandardized and (b) standardized deviance residuals.*

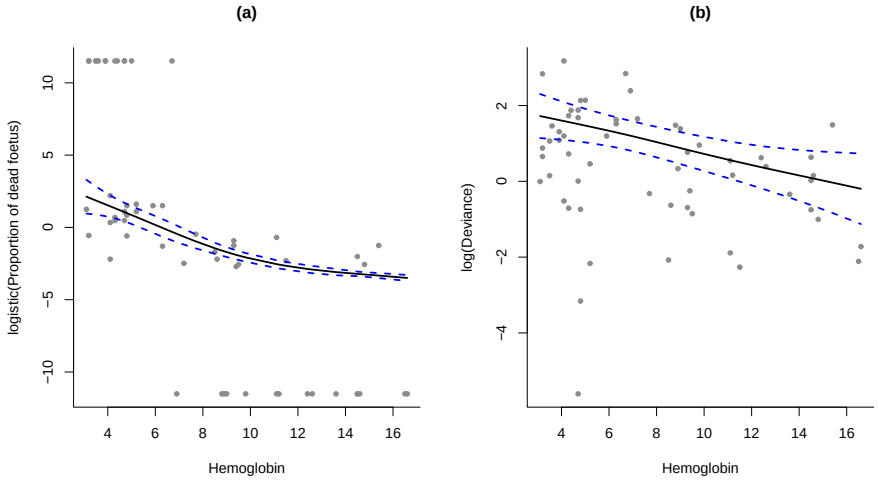


Figure 5.10: *The low iron rat data. The estimated functions (a) $\eta(\cdot)$ and (b) $\xi(\cdot)$ (solid curves) with approximate confidence intervals (dashed curves).*

5.5 The Lidar data

An example of normal data is the Lidar dataset, which is discussed in Ruppert *et al.* (2003), among others. Lidar (Light detection and ranging) is a technique to detect chemical compounds in the atmosphere. The dataset contains the data coming from a Lidar experiment: 221 observations of the distance the light traveled before being reflected back to its source, and of the logarithm of the ratio of received light from two laser sources. In Figure 5.11 we plot the data and the obtained estimates for the mean and the variance function. The data show indeed a non-constant variance and the estimate seems to catch the increasing shape of the function. Figure 5.12 gives a clear indication that the esti-

mation of the variance function performed quite well. Ruppert *et al.* (2003) also find an increasing shape for the variance function of these data.

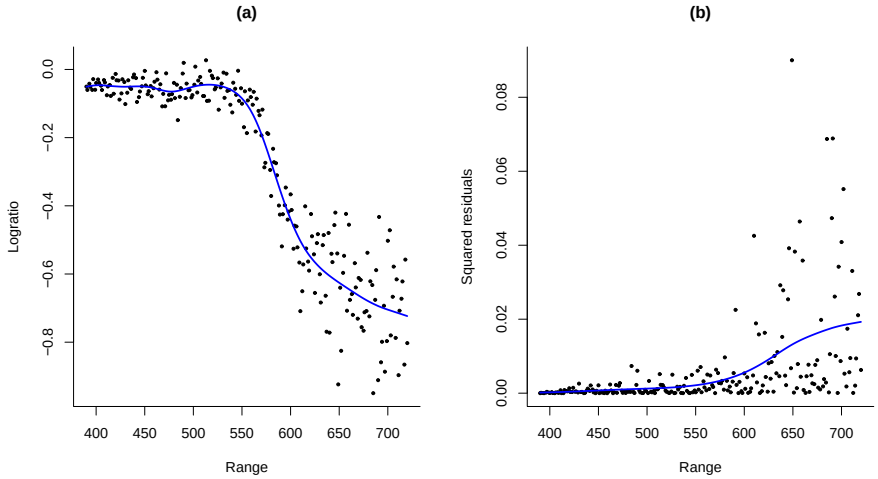


Figure 5.11: *The Lidar data: (a) mean and (b) variance function estimation* ($k_\mu = 45$ $k_\gamma = 35$, $p_\mu = p_\gamma = 3$*).*

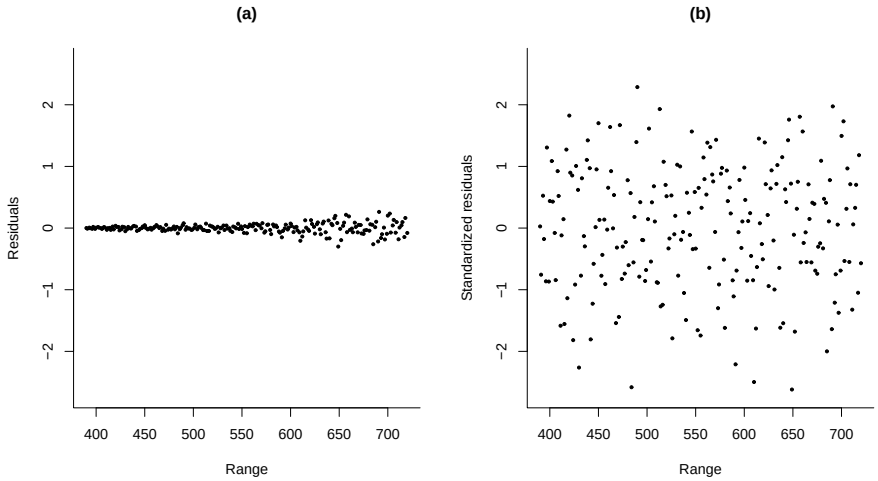


Figure 5.12: *The Lidar data: (a) unstandardized and (b) standardized residuals.*

Figures 5.13 (a) and (b) show the estimated functions $\eta(\cdot)$ and $\xi(\cdot)$ respectively, together with their approximate confidence intervals.

6 Further discussion

As was noted in Lee and Nelder (2000), the double exponential family leads to identical inference as the extended quasi likelihood approach (see McCullagh and Nelder (1989)).

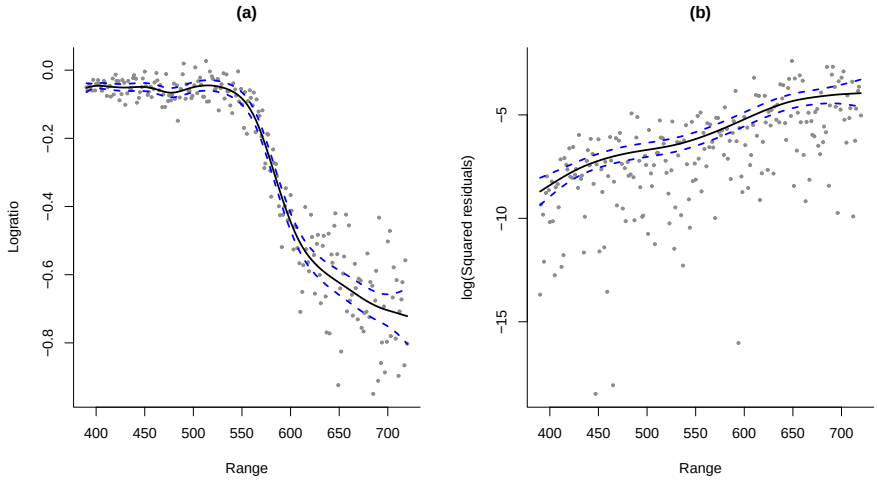


Figure 5.13: *The Lidar data. The estimated functions (a) $\eta(\cdot)$ and (b) $\xi(\cdot)$ (solid curves) with approximate confidence intervals (dashed curves).*

In a similar manner one can consider an extended quasi likelihood approach in the current context. This has been done by the authors but for brevity we do not elaborate on this in the paper.

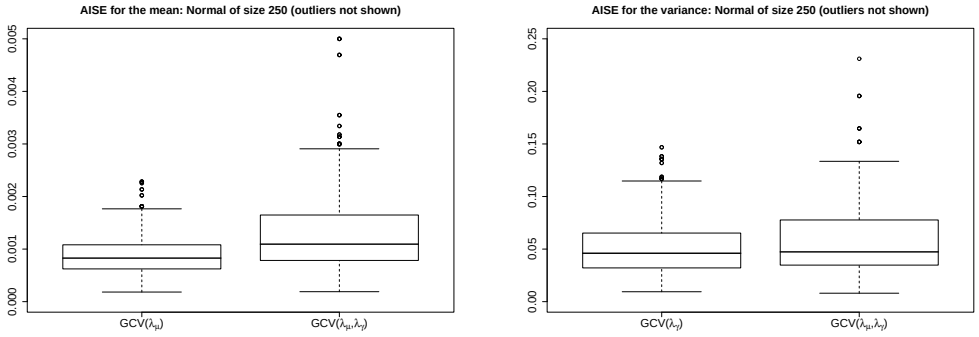
One might wonder whether it is possible to select the two smoothing parameters λ_μ and λ_γ simultaneously. A first task then is to define an appropriate type of bivariate GCV criterion. Following the idea behind the univariate GCV criteria developed in Section 3 one needs to define the deviance of the model: start from the likelihood in (3.1) and take $\theta_S = g^{-1}(y)$ and $\gamma_S = d(y, \theta)$ to be the choice of θ and γ corresponding to the saturated model. As the total number of degrees of freedom we take the sum of the equivalent number of degrees of freedom needed when fitting the mean and the dispersion function. This leads to

$$\mathrm{GCV}(\lambda_\mu, \lambda_\gamma) = \frac{\sum_{i=1}^{n} \left(\log \hat{\gamma}_{\lambda_\gamma}(x_i) - \log d(y_i, \hat{\theta}_{\lambda_\mu}(x_i)) + d(y_i, \hat{\theta}_{\lambda_\mu}(x_i))/\hat{\gamma}_{\lambda_\gamma}(x_i) \right)}{(n - \mathrm{df}(\lambda_\gamma) - \mathrm{df}(\lambda_\mu))^2}. \quad (6.1)$$

Minimizing $\mathrm{GCV}(\lambda_\mu, \lambda_\gamma)$ with respect to λ_μ and λ_γ simultaneously will then give direct choices for both smoothing parameters. The optimization can be done numerically using the `optim` function of R. Experiences show that optimization of $\mathrm{GCV}(\lambda_\mu, \lambda_\gamma)$ is not straightforward: the function often presents local minima in different directions, so that the minimization process becomes difficult and strongly dependent on the provided initial values. Alternatively, one can do a two-dimensional grid search, but this at the cost of a considerable increase in computing time. A grid search also faces the same problem of

several local minima.

As an illustration we present in Figure 6.1 boxplots of the AISE values for the mean and the dispersion estimation in the normal model when choosing the smoothing parameters either in separate steps (left-hand side boxplots) or via the bivariate criterion (right-hand side boxplots). As can be seen there is no gain from considering the bivariate GCV criterion. Similar results (not presented here) were obtained for the other simulation models. Since the bivariate GCV criterion also seems to suffer from the above mentioned drawbacks, we do not recommend its use.



(a) (b)

Figure 6.1: *Normal model with $n = 250$: boxplots of the AISE-values for the mean (a) and the variance (b) estimation when choosing the smoothing parameters in two different steps (left boxplots) or simultaneously (right boxplots).*

Other selection criteria for choosing λ_μ and λ_γ have been explored, among which Akaike's Information (AIC) criteria (univariate and bivariate). None of these however outperformed the univariate GCV criteria of Section 3. Nevertheless, these criteria are not without problems since the optimization problems can be hard, in particular with respect to the smoothing parameter selection for the dispersion estimation. More research is needed for developing better selection rules.

In this paper the mean and dispersion function are estimated nonparametrically using P-spline techniques. Other smoothing techniques such as local polynomial fitting could also serve as a basis here. So far no theoretical studies have been done for establishing the rate of convergence of the dispersion function estimator. Concerning estimation of the mean regression function using penalized splines techniques, there is recent work dealing with asymptotic theory. This includes work by Li and Ruppert (2008) who establish the asymptotic distribution of penalized mean regression estimators. Further, Kauermann *et al.* (2009) consider generalized penalized spline smoothing, and adopt a mixed-model

approach and Laplace approximations for their asymptotic study. Claeskens *et al.* (2009) study asymptotic properties of a class of penalized spline regression estimators. Asymptotic results focusing on the choice of the difference type of penalty are provided in Gijbels and Verhasselt (2010a,b). Some studies on rates of convergence for variance function estimation (in general) can be found in the literature, including the works by Hall and Carroll (1989), Ruppert *et al.* (1997), Fan and Yao (1998) and Wang *et al.* (2008). These papers in particular study the effect of mean estimation on variance estimation, and reveal, for example, that the rate of convergence of the variance estimator is similar to that of the mean estimator provided the mean function has some minimal degree of smoothness. It would be interesting to investigate the effect of the mean estimation on the dispersion function estimation, and to establish asymptotic theory for the dispersion function estimate.

In Section 4 we provide approximate confidence intervals. Of interest would also be to develop a method for constructing confidence bands. This is an open research issue.

Acknowledgements

The authors thank the reviewers for their valuable comments which led to additional interesting investigations and an improved presentation. Support from the GOA/07/04-project of the Research Fund KULeuven is gratefully acknowledged, as well as support from the IAP research network nr. P6/03 of the Federal Science Policy, Belgium. The authors thank Dr Danie Uys for helpful discussions.

References

Aerts, M. and Claeskens, G. (1997). Local polynomial estimation in multiparameter likelihood models. *Journal of the American Statistical Association*, **92**, 1536–1545.

Alonso, J., Muñoz, A. and Antó, J. M. (1996). Using length of stay and inactive days in the hospital to assess appropriateness of utilization in Barcelona, Spain. *Journal of Epidemiology and Community Health*, **50**, 196–201.

de Boor, C. (2001). *A Practical Guide to Splines*. Revised Edition. Springer, Berlin.

Carroll, R.J. and Ruppert, D. (1988). Transformation and Weighting in Regression, Chapman and Hall: New York.

Claeskens, G., Krivobokova, T. and Opsomer, J.D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, **96**, 529–544.

Davidian, M. and Carroll, R.J. (1987). Variance function Estimation. *Journal of the American Statistical Association*, **82**, 1079–1091.

Davidian, M. and Carroll, R.J. (1988). A note on Extended Quasi-likelihood. *Journal of the Royal Statistical Society, Series B*, **50**, 74–82.

Dey, D.K., Galfand, A.E. and Peng, F. (1997). Overdispersed generalized linear models. *Journal of Statistical Planning and Inference*, **64**, 93–107.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11**, 89–121.

Efron, B. (1986). Double Exponential Families and their Use in Generalized Linear Regression. *Journal of the American Statistical Association*, **81**, 809–721.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645–660.

Galfand, A.E. and Dalal, S.R. (1990). A note on Overdispersed Exponential Families. *Biometrika*, **77**, 55–64.

Gijbels, I. and Verhasselt, A. (2010a). P-splines regression smoothing and difference type of penalty. *Statistics and Computing*, to appear.
DOI:10.1007/s11222-009-9140-0

Gijbels, I. and Verhasselt, A. (2010b). Regularisation and P-splines in generalised linear models. *Journal of Nonparametric Statistics*, to appear.
DOI: 10.1080/10485250903365900

Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society, Series B*, **51**, 3–14.

Hall, P., Kay, J. and Titterington, D. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.

Hinde, J. (1982). Compound Poisson regression mode: in *GLIM* 82; In *Proceedings of the International Conference on Generalized Linear Models*, ed. Gilchrist, R., 109–121, Springer: New York.

Hinde, J. and Demétrio C.G.B. (1998). Overdispersion: Models and estimation. *Computational Statistics & Data Analysis*, **27**, 151–170.

Kauermann, G. , Krivobokava, T. and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B*, **71**, 487–503.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee, Y. and Nelder, J. A. (2000). The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ratio data. *Applied Statistics*, **49**, 413–419.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, **95**, 415–436.

Marx, B.D and Eilers, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28**, 193–209.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Model*, Chapman and Hall: London.

Moore, D. F. and Tsiatis, A. (1991). Robust estimation of the variance in Moment Methods for Extra-Binomial and Extra-Poisson Variation. *Biometrics*, **47**, 383–401.

Nelder, J.A. and Lee, Y. (1992). Likelihood, Quasi-likelihood and Pseudolikelihood: Some Comparisons. *Journal of the Royal Statistical Society, Series B*, **54**, 273–284.

Nelder, J.A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

Nott, D. (2006). Semiparametric estimation of mean and variance functions for non-Gaussian data. *Computational Statistics*, **21** 603–620.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507–554.

Ruppert, D., Wand, M. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Ruppert, D., Wand, M.P. Holst, U. and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, **39**, 262–273.

Stasinopoulos, D. M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, Issue 7.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wang, L., Brown, L.D., Cai, T.T. and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression *The Annals of Statistics*, **36**, 646–664.